(12) UK Patent Application (19) GB (11) 2 212 636 (13) A
(43) Date of A publication 26.07.1989

(21) Application No 8826207.6

(22) Date of filing 09.11.1988

(30) Priority data
(31) 121500 (32) 17.11.1987 (33) US

(71) Applicant
Amoco Corporation
(Incorporated in the USA - Indiana)
200 East Randolph Drive, Chicago, Illinois 60601, United States of America

(72) Inventors
Delbert Carl Johnson
Lawrence Lamar Sorensen
Susan Pyeatt Ennis

(74) Agent and/or Address for Service
Mathys and Squire
10 Fleet Street, London, EC4Y 1AY, United Kingdom

(51) INT CL⁴
G06F 15/40, G01V 1/28

(52) UK CL (Edition J)
G4A AUB
U1S S2141

(56) Documents cited
GB 2182796 A EP 0217655 A EP 0157354 A
EP 0144202 A EP 0062777 A

(58) Field of search
UK CL (Edition J) G4A AUA AUB, G4R REX
INT CL⁴ G06F

(54) Identifying data format

(57) A method is disclosed of identifying, from a list of known data formats, a particular format for a set of data. A representation of at least a portion of the data is created and from the representation, characteristics of the data are obtained. Utilizing predetermined logic rules, the data characteristics are matched to known data characteristics of each known data format until a match is accomplished. Thereafter, an indication is generated of the data format that has been matched.

GB 2 212 636 A

"METHOD FOR DETERMINING THE FORMAT OF SEISMIC TRACE DATA"

The present invention relates to a method for identifying the format of seismic trace data and, more particularly, to such a method which uses internal logic rules, such as provided by artificial intelligence, to identify the format.

The successful processing of data is highly dependent upon the particular format in which the data has been recorded. For example, companies that process seismic traces acquire the data from many different service companies, each of which records the data in their own particular format(s). In order for the processing to be successfully accomplished, the data needs to be placed into a format that is compatible with the particular computer programs that are to be used to process the data, for example programs to increase the signal-to-noise ratio. Various data recording standards have been published by organizations; however, these standards have not been widely adhered to so seismic data processing includes the need for an initial determination of the format of the data that has been received from an outside source.

Most companies that process data have developed or acquired a suite of data processing programs that process the data in the many different formats, i.e., a single processing program for each format or set of

related formats. Again, the problem is to select the appropriate processing program for the particular format, as well as the determination of certain peculiar processing parameters that are to be used in the processing. Therefore, there is a need for a simple method of determining the format of the data.

The present invention has been contemplated to overcome the forgoing deficiencies and to meet the above-described need. The present invention provides a method of identifying, from a list of known data formats, a particular format for a set of data. In the method, a representation is created of at least a portion of the data, and from that representation, characteristics of the data are obtained. Utilizing predetermined logic rules, supplied by experienced users and translated into a form used by an expert system shell, the data characteristics are matched to known data characteristics for each known data format until a match is accomplished. Expert system shells are programs written for programmable digital computers that manipulate symbols in predefined ways. In particular, they allow the backward chaining of the detailed description. Thereafter, a report is generated indicating that the data is in a particular matched data format. By using this program, a set of data can be easily and quickly reviewed and an indication of the particular data format is provided, as well as any particular additional known parameters that are required in the processing of the data.

--

The present invention provides a method of identifying, from a list of known data formats, a particular format for a set of data by using a programmable digital computer. While this method can be utilized for any data format, for the purpose of this discussion, the field of use will be identified as processing seismic data used in the exploration for oil and gas. Basically, once a seismic tape containing data is acquired from an outside

source, it is first loaded on a computer system and a representation of the content of the data is created. This representation is a uniform formatted representation, i.e., a formatted hexidecimal dump of the data. An example of a representation is shown in Table I.

The representation is electronically transmitted to the computer where programs of the present invention reside. When the user invokes the program, the name of the file containing the representation must be supplied to the program and other particular information (described herein below) about the file that might be available may be supplied. With the available information, the program then scans the first portion of the representation to identify the location and value of particular data items to obtain characteristics of the data. Then, the program identifies the particular data format, the name of the particular processing program useful in processing data in that format and the values of any parameters that might be needed during such processing.

Specifically, before the program of the present invention can be run, it is necessary to transform the data set into a uniformally formatted representation. The first portion, for example 480 bytes, of each physical record of the tape are generated in hexadecimal form. Each record is then represented by a line that contains the number of bytes in the physical record and the number of the physical record, followed by 12 lines containing the hexadecimal representation of the data. The hexadecimal representation is arranged in groups of four nibbles separated by a single space.

Once this has been transferred to the machine where the program runs, the following information is asked of the user:

1. The name of the file containing the input data.

2. Whether observers' notes are available for this data, and if yes:

(a) the number of traces per seismic record.

(b) the number of auxiliary traces per seismic record.

(c) the sample interval used for recording the data.

(d) the length of the trace data in seconds.

3. Whether a hardcopy of the representation is available to user and if so the format used to record the trace data.

4. Whether the data was initially recorded in analog or digital form.

The representation of the data is then analyzed to obtain data characteristics, for example, the program will scan the data representation at particular locations and retrieve the values at such locations. These characteristics are thereafter used in the matching process with logic rules. The logic rules utilized in the present invention are a set of knowledge relationships captured from experienced people and written using the Personal Consultant Plus©, a computer program marketed by Texas Instruments Corporation. The program accomplishes its determination by the application of the set of rules that have been determined to solve this problem correctly about 80-85% of the time. Each of the rules is an independent piece of information that is known to experienced people who can make a parameter determination. The Personal Consultant Plus© determines the manner and order by which these rules are used in any particular run of the program.

Some of the logical rules can be classified as facts about the problem of program parameter determination. Some of the rules can be classified as generally disseminated practices about the problem of recognizing seismic formats. A third classification of rules is that one developed after long discussions with experienced people who are familiar with the program and parameter determination and truly reflects their expertise in per-

forming these tasks. Table II shows examples of these roles.

One of the strengths of this program is that there can be any number of rules that have been accumulated and coded into the system, thereby permitting a high degree of confidence in the outputted format determination.

The predetermined logic rules developed are applied against the characteristics of the data by a logic process called backward chaining. In other words, a first or trial data format is chosen from the list of known data formats and set as a goal. For such data to be in the first data format, one or several rules must be activated, i.e., have all of their premises be proved true. The required rules are found and their premises are examined. If the facts needed to determine the truth of the premises are not known, then these facts are set as subgoals, and the cycle of selecting rules occurs again. If a sufficient set of rules is able to activate, there is a match for the rules, then the data is in the first data format. However, if the rules do not activate, then a new trial data format is selected as a goal, the respective logic rules are found, and the backward chaining of rules and facts is applied as before. The program will continue with each known data format until a match is found. If no match is found, then the program will indicate that no match was found.

When the program has determined to the predetermined satisfaction limit the particular format of the data, an indication or display is provided to the user of the name of the preferred processing program and any associated processing variables, needed to process the data. Six examples of the indication are provided in Table III.

Wherein the present invention has been described in particular relation to the examples included herein, it should be understood that other and further modifications, apart from those shown or suggested herein, may be made within the scope and spirit of the present invention.

TABLE 1

BYTES READ 3200 RECORD 1

1	C3D3	C9C5	D5E3	40C1	D4D6	C3D6	40C1	D9C5	C140	E3F3	C161	F540	D3C9	D5C5	40F1	F7F5	4040	4040	4040	4040
41	4040	4040	4040	4040	4040	4040	4040	4040	4040	4040	4040	4040	4040	4040	4040	4040	4040	4040	4040	4040
81	C4C5	D4E4	E740	E2C5	C740	4040	40F1	F7F5	E240	C6D6	D940	E2D7	7DE2	40F1	6160	F2F7	F0F0	4040	4040	4040
121	4040	4040	4040	4040	4040	4040	4040	4040	4040	4040	4040	4040	4040	4040	4040	4040	4040	4040	4040	4040
161	C4C5	D5E2	C9E3	E840	F6F2	F5F0	F7F9	F4F8	C340	4040	4040	4040	4040	4040	4040	4040	4040	4040	4040	4040
201	4040	4040	4040	4040	4040	4040	4040	4040	4040	4040	4040	4040	4040	4040	4040	4040	4040	4040	4040	4040
241	C5D5	C440	C5C2	C3C4	C9C3	4040	4040	4040	4040	4040	4040	4040	4040	4040	4040	4040	4040	4040	4040	4040

BYTES READ 400 RECORD 2

1	0000	0000	0000	00AF	0000	1F0C	0010	0006	07D0	07D0	0FA0	0000	0001	0078	0001	0001	0000	0000	0000	0000
41	0000	0000	0000	0000	0001	0000	0000	0002	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000
81	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000
121	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000
161	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000
201	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000
241	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000

BYTES READ 16240 RECORD 3

1	0000	96F1	0000	0001	0000	01BA	0000	0001	0000	0000	0000	01BA	0000	0001	0001	0000	0000	0001	0000	0001
41	0000	002B	0000	0000	0000	0051	FFFF	FFAF	0000	0000	0000	0000	0000	0000	0001	0001	0005	2080	0001	A130
81	0000	0000	0000	0000	0001	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0FA0	07D0	0000
121	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000
161	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000
201	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000
241	BF19	C580	BE4A	2800	3EFA	6800	3F12	3300	3F10	9B00	3EEF	8800	3E99	B000	BE35	6000	BF15	A480	BF22	CE00

BYTES READ 16240 RECORD 4

1	0000	96F2	0000	0002	0000	01BA	0000	0002	0000	0000	0000	01BA	0000	0002	0001	0000	0000	0001	0000	0002
41	0000	001E	0000	0000	0000	0051	FFFF	FFAF	0000	0000	0000	0000	0000	0000	0001	0001	0005	2080	0001	A130
81	0000	0000	0000	0000	0001	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0FA0	07D0	0000
121	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000
161	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000
201	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000
241	BF35	8400	BF37	4000	BF30	6E00	BF2C	5E00	BF2D	AE00	BF2C	3800	BF22	9600	BF1A	8800	BEF2	5800	BEA8	2800

TABLE II

Rule069 STD-RULES/antecedent

If 1) the sample interval used to record the data in microseconds is known, and

2) the measure of certainty associated with the sample interval used to record the data in microseconds,

Then, it is definite (100%) that the sample interval used to record the data in milliseconds is the sample interval used to record the data in microseconds divided by 1000.

```
IF:   SI IS KNOWN AND CERTAINTY SI
THEN: SI-MILLI = VALUE SI / 1000

PREMISE:  ($AND
            (KNOWN FRAME SI)
            (MEASURE1 FRAME SI))
ACTION:   (DO-ALL
            (CONCLUDE FRAME SI-MILLI
              (FQUOTIENT
                (VAL1 FRAM SI) 1000) TALLY 100))

ANTECEDENT: YES
```

Rule074 STD-RULES

If 1) both I*4 and R*4 are equally likely, and

2) the measure of certainty associated with the format of the recorded data,

Then there is weakly suggestive evidence (20%) that the format of the recorded data is R*4.

```
IF:  SEL-DFMT (VAL FRAME DFMT) AND CERTAINTY DFMT
THEN:  DFMT = R*4 CF 20

PREMISE:  ($AND
            (SEL-DFMT
              (VAL FRAME DFMT))
            (MEASURE1 FRAME DFMT))
ACTION:   (DO-ALL
            (CONCLUDE FRAME DFMT R*4 TALLY 20))
```

Rule072 STD-RULES

If the number of line headers is 0,

Then, 1) there is strongly suggestive evidence (80%) that the first 32 words of the binary header is all zeros, and

2) there is strongly suggestive evidence (80%) that the tape is a variant of the SEG-Y format.

```
IF:  NO-LH-1 = 0
THEN:  VALUE-BH = GET-0-BH CF 80 AND A-SEGY CF 80

PREMISE:  ($AND
            (SAME FRAME NO-LH-1 0))
ACTION:   (DO-ALL
            (CONCLUDE FRAME VALUE-BH
              (GET-0-BH) TALLY 80)
            (CONCLUDE FRAME A-SEGY YES TALLY 80))
```

TABLE III

Example 1. The system found that there were only seven seconds of data in the file even though the user had indicated that there were eight.

MPS-STD-1 CONCLUSIONS:
A major recommendation is as follows: Use the program EXCH to reformat the file, fully accounting for all of the subsidiary recommendations. (74%)

The complete list of parameters for the selected program is as follows:

Sample interval		2	97%
Number of regular traces		16	87%
Number of auxiliary traces		0	87%
Number of samples per trace		3500	74%
Length of the trace header		240	74%
Number of line headers		2	90%
Data format		R*4	100%
Record number position	bytes 11 and 12		100%
Trace number position	bytes 15 and 16		100%

A subsidiary recommendation is as follows: It appears that the recording has been shortened to 7 seconds. (80%)

Example 2. The system notes that the record numbers are large at the beginning of the file and might exceed the program capability by the time the end of the file is reached.

MPS-STD-1 CONCLUSIONS:
A major recommendation is as follows: Use the program EXCH to reformat the file, fully accounting for all of the subsidiary recommendations. (83%)

The complete list of parameters for the selected program is as follows:

Sample interval	2	97%
Number of regular traces	48	88%
Number of auxiliary traces	0	99%
Number of samples per trace	3500	83%
Length of the trace header	240	83%
Number of line headers	2	90%
Data format	R*4	100%
Record number position	bytes 11 and 12	100%
Trace number position	bytes 15 and 16	100%

A subsidiary recommendation is as follows: The data contains record numbers that are greater than 16,000. The file can be reformatted, but the records should be renumbered (72%)

Example 3. The system notes that there is an additional tenth of a second of data in the file.

MPS-STD-1 CONCLUSIONS:
A major recommendation is as follows: Use the program EXCH to reformat the file, fully accounting for all of the subsidiary recommendations. (40%)

The complete list of parameters for the selected program is as follows:

Sample interval		4	95%
Number of regular traces		24	40%
Number of auxiliary traces		0	53%
Number of samples per trace		1175	65%
Length of the trace header		240	65%
Number of line headers		2	90%
Data format		R*4	100%
Record number position	bytes 221 and 222		80%
Trace number position	bytes 223 and 224		80%

A subsidiary recommendation is as follows: It appears that there are actually 4.7 seconds of data, rather than the 4.6 seconds indicated in the observer notes. This file appears to be in CDP sort sequence. (72%)

Example 4. For the case of a SEGD formatted file, additional parameters are not required to process the file.

M-STD-1 CONCLUSIONS:
A major recommendation is as follows: The program SEGD should be used to reformat the file because it is in SEG-D format.

Example 5. Here two problems were found. The record numbers were not recorded and the traces have been shortened.

MPS-STD-1 CONCLUSIONS:
A major recommendation is as follows: Use the program EXCH to reformat the file, fully accounting for all of the subsidiary recommendations. (43%)

The complete list of parameters for the selected program is as follows:

Sample interval	2	97%
Number of regular traces	48	54%
Number of auxiliary traces	0	86%
Number of samples per trace	1000	94%
Length of the trace header	240	94%
Number of line headers	2	90%
Data format	R*4	100%
Record number position	unknown	43%
Trace number position	bytes 15 and 16	54%

A subsidiary recommendation is as follows: It appears that the recording has been shortened to 2 seconds. (80%) There are no record numbers in the trace headers. The file can be processed, but the record numbers will need to be generated by renumbering. (43%)

Example 6. There may be a problem with the interpretation of this case because the user didn't dump enough of the seismic data file as input to the program.

MPS-STD-1 CONCLUSIONS:
A major recommendation is as follows: Use the program EXCH to reformat the file, fully accounting for all of the subsidiary recommendations. (43%)

The complete list of parameters for the selected program is as follows:

Sample interval	2	97%
Number of regular traces	18	54%
Number of auxiliary traces	0	86%
Number of samples per trace	3000	92%
Length of the trace header	240	92%
Number of line headers	2	90%
Data format	R*4	93%
Record number position	unknown	43%
Trace number position	bytes 15 and 16	54%

A subsidiary recommendation is as follows: The number of traces MAY be limited by the number of records that have been dumped, you ought to dump more records and re-run this consultation. Both the binary header and the observer notes indicate that there are 48 traces.

CLAIMS

1. A method of identifying, from a list of known data formats, the particular format for a set of data, comprising:

(a) creating a representation of at least a portion of the data;

(b) from the representation, obtaining characteristics of the data;

(c) utilizing predetermined logic rules, matching the data characteristics of (b) to known data characteristics for the known data formats until a match is accomplished; and

(d) generating an indication that the data is in the matched data format.

2. The method of Claim 1 wherein the representation of the data is uniformly formatted.

3. The method of Claim 2 wherein the uniformly formatted data is in hexidecimal form.

4. The method of Claim 1 wherein step (b) comprises analyzing a first portion of the data to identify the location and value of particular data items.

5. The method of Claim 1 wherein before step (c), including the step of inputting user-known data characteristics.

6. The method of Claim 1 wherein step (d) includes utilizing known data processing requirements for each known data format, generating an indication of preferred processing variables.

7. The method of Claim 1 wherein the set of data are seismic data traces.